



ELHnet: a convolutional neural network for classifying cochlear endolymphatic hydrops imaged with optical coherence tomography

GEORGE S. LIU,¹ MICHAEL H. ZHU,² JINKYUNG KIM,¹ PATRICK RAPHAEL,¹
BRIAN E. APPLGATE,³ AND JOHN S. OGHALAI^{4,*}

¹Department of Otolaryngology–Head and Neck Surgery, Stanford University, 801 Welch Road, Stanford, CA 94305, USA

²Department of Computer Science, Stanford University, 353 Serra Mall, Stanford, CA 94305, USA

³Department of Biomedical Engineering, Texas A&M University, 5059 Emerging Technology Building, 3120 TAMU, College Station, TX 77843, USA

⁴USC Caruso Department of Otolaryngology–Head and Neck Surgery, 1540 Alcazar, Suite 204M, Los Angeles, CA 90033, USA

*oghalai@usc.edu

Abstract: Detection of endolymphatic hydrops is important for diagnosing Meniere’s disease, and can be performed non-invasively using optical coherence tomography (OCT) in animal models as well as potentially in the clinic. Here, we developed ELHnet, a convolutional neural network to classify endolymphatic hydrops in a mouse model using learned features from OCT images of mice cochleae. We trained ELHnet on 2159 training and validation images from 17 mice, using only the image pixels and observer-determined labels of endolymphatic hydrops as the inputs. We tested ELHnet on 37 images from 37 mice that were previously not used, and found that the neural network correctly classified 34 of the 37 mice. This demonstrates an improvement in performance from previous work on computer-aided classification of endolymphatic hydrops. To the best of our knowledge, this is the first deep CNN designed for endolymphatic hydrops classification.

© 2017 Optical Society of America

OCIS codes: (100.4996) Pattern recognition, neural networks; (170.0170) Medical optics and biotechnology; (170.4500) Optical coherence tomography.

References and links

1. R. G. Chelu, K. W. Wanambiro, A. Hsiao, L. E. Swart, T. Voogd, A. T. van den Hoven, M. van Kranenburg, A. Coenen, S. Boccalini, P. A. Wielopolski, M. W. Vogel, G. P. Krestin, S. S. Vasanaawala, R. P. J. Budde, J. W. Roos-Hesselink, and K. Nieman, “Cloud-processed 4D CMR flow imaging for pulmonary flow quantification,” *Eur. J. Radiol.* **85**(10), 1849–1856 (2016).
2. K. Kamnitsas, C. Baumgartner, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, A. Nori, A. Criminisi, D. Rueckert, and B. Glocker, “Unsupervised domain adaptation in brain lesion segmentation with adversarial networks,” *Inf. Process. Med. Imaging* (2016).
3. S. Ö. Arik, B. Ibragimov, and L. Xing, “Fully automated quantitative cephalometry using convolutional neural networks,” *J. Med. Imaging (Bellingham)* **4**(1), 014501 (2017).
4. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature* **542**(7639), 115–118 (2017).
5. M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, “Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network,” *IEEE Trans. Med. Imaging* **35**(5), 1207–1216 (2016).
6. K. Lekadir, A. Galimzianova, A. Betriu, M. del M. Vila, L. Igual, D. Rubin, E. Fernandez, P. Radeva, and S. Napel, “A Convolutional Neural Network for Automatic Characterization of Plaque Composition in Carotid Ultrasound,” *IEEE J. Biomed. Health Inform.* **2194**, 48–55 (2016).
7. F. Pereira, A. Bueno, A. Rodriguez, D. Perrin, G. Marx, M. Cardinale, I. Salgo, and P. Del Nido, “Automated detection of coarctation of aorta in neonates from two-dimensional echocardiograms,” *J. Med. Imaging (Bellingham)* **4**(1), 014502 (2017).
8. G. G. Gardner, D. Keating, T. H. Williamson, and A. T. Elliott, “Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool,” *Br. J. Ophthalmol.* **80**(11), 940–944 (1996).
9. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Adv. Neural Inf. Process. Syst.* **2012**, 1097–1105 (2012).

10. P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks," (2013).
11. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2015), **07–12**–June, pp. 1–9.
12. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2016), pp. 770–778.
13. N. H. Cho, J. H. Jang, W. Jung, and J. Kim, "In vivo imaging of middle-ear and inner-ear microstructures of a mouse guided by SD-OCT combined with a surgical microscope," *Opt. Express* **22**(8), 8985–8995 (2014).
14. H. M. Subhash, V. Davila, H. Sun, A. T. Nguyen-Huynh, A. L. Nuttall, and R. K. Wang, "Volumetric in vivo imaging of intracochlear microstructures in mice by high-speed spectral domain optical coherence tomography," *J. Biomed. Opt.* **15**(3), 036024 (2010).
15. H. Y. Lee, P. D. Raphael, J. Park, A. K. Ellerbee, B. E. Applegate, and J. S. Oghalai, "Noninvasive in vivo imaging reveals differences between tectorial membrane and basilar membrane traveling waves in the mouse cochlea," *Proc. Natl. Acad. Sci. U.S.A.* **112**(10), 3128–3133 (2015).
16. H. Y. Lee, P. D. Raphael, A. Xia, J. Kim, N. Grillet, B. E. Applegate, A. K. Ellerbee Bowden, and J. S. Oghalai, "Two-Dimensional Cochlear Micromechanics Measured In Vivo Demonstrate Radial Tuning within the Mouse Organ of Corti," *J. Neurosci.* **36**(31), 8160–8173 (2016).
17. A. Xia, X. Liu, P. D. Raphael, B. E. Applegate, and J. S. Oghalai, "Hair cell force generation does not amplify or tune vibrations within the chicken basilar papilla," *Nat. Commun.* **7**, 13133 (2016).
18. S. S. Gao, A. Xia, T. Yuan, P. D. Raphael, R. L. Shelton, B. E. Applegate, and J. S. Oghalai, "Quantitative imaging of cochlear soft tissues in wild-type and hearing-impaired transgenic mice by spectral domain optical coherence tomography," *Opt. Express* **19**(16), 15415–15428 (2011).
19. S. D. Rauch, S. N. Merchant, and B. A. Thedinger, "Meniere's syndrome and endolymphatic hydrops. double-blind temporal bone study," *Ann. Otol. Rhinol. Laryngol.* **98**(11), 873–883 (1989).
20. A. N. Salt and S. K. Plontke, "Endolymphatic Hydrops: Pathophysiology and Experimental Models," *Otolaryngol. Clin. North Am.* **43**(5), 971–983 (2010).
21. J. Kim, X. Liu, Z. Jawadi, N. Grillet, and J. Oghalai, "Acute changes in the mouse cochlea after blast injury," in *Abstracts of the Midwinter Research Meeting of the Association for Research in Otolaryngology 2016*. (2016).
22. F. Fiorino, F. B. Pizzini, A. Beltramello, and F. Barbieri, "MRI performed after intratympanic gadolinium administration in patients with Ménière's disease: Correlation with symptoms and signs," *Eur. Arch. Otorhinolaryngol.* **268**(2), 181–187 (2011).
23. H. Fukuoaka, Y. Takumi, K. Tsukada, M. Miyagawa, T. Oguchi, H. Ueda, M. Kadoya, and S. Usami, "Comparison of the diagnostic value of 3 T MRI after intratympanic injection of GBCA, electrocochleography, and the glycerol test in patients with Meniere's disease," *Acta Otolaryngol.* **132**(2), 141–145 (2012).
24. H. F. Schuknecht, *Pathology of the Ear*, 2nd ed. (Lea & Febiger, 1993).
25. S. I. Cho, S. S. Gao, A. Xia, R. Wang, F. T. Salles, P. D. Raphael, H. Abaya, J. Wachtel, J. Baek, D. Jacobs, M. N. Rasband, and J. S. Oghalai, "Mechanisms of Hearing Loss after Blast Injury to the Ear," *PLoS One* **8**(7), e67618 (2013).
26. A. N. Salt and S. K. Plontke, "Endolymphatic Hydrops: Pathophysiology and Experimental Models," *Otolaryngol. Clin. North Am.* **43**(5), 971–983 (2010).
27. S. F. Klis, J. Buijs, and G. F. Smoorenburg, "Quantification of the relation between electrophysiologic and morphologic changes in experimental endolymphatic hydrops," *Ann. Otol. Rhinol. Laryngol.* **99**(7), 566–570 (1990).
28. P. Dollar, Z. Tu, H. Tao, and S. Belongie, "Feature mining for image classification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2007).
29. G. S. Liu, J. Kim, B. E. Applegate, and J. S. Oghalai, "Computer-aided detection and quantification of endolymphatic hydrops within the mouse cochlea in vivo using optical coherence tomography," *J. Biomed. Opt.* **22**(7), 076002 (2017).
30. L. Fang, D. Cunefare, C. Wang, R. H. Guymier, S. Li, and S. Farsiu, "Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search," *Biomed. Opt. Express* **8**(5), 2732–2744 (2017).
31. F. G. Venhuizen, B. van Ginneken, B. Liefers, M. J. J. P. van Grinsven, S. Fauser, C. Hoyng, T. Theelen, and C. I. Sánchez, "Robust total retina thickness segmentation in optical coherence tomography images using convolutional neural networks," *Biomed. Opt. Express* **8**(7), 3292–3316 (2017).
32. S. P. K. Karri, D. Chakraborty, and J. Chatterjee, "Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration," *Biomed. Opt. Express* **8**(2), 579–592 (2017).
33. S. Gao, P. D. Raphael, R. Wang, J. Park, A. Xia, B. E. Applegate, and J. S. Oghalai, "In vivo vibrometry inside the apex of the mouse cochlea using spectral domain optical coherence tomography," *Biomed. Opt. Express* **4**(2), 230–240 (2013).
34. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.* **115**(3), 211–252 (2015).
35. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Int.*

- Conf. Learn. Represent. 1–14 (2015).
36. A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, “EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos,” *IEEE Trans. Med. Imaging* **36**(1), 86–97 (2017).
 37. D. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *Int. Conf. Learn. Represent.* 1–13 (2014).
 38. Theano Development Team, “Theano: A Python framework for fast computation of mathematical expressions,” arXiv e-prints **abs/1605.0**, (2016).
 39. F. Chollet, “Keras,” <https://github.com/fchollet/keras>.
 40. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
 41. M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks BT - Computer Vision – ECCV 2014,” in *Computer Vision – ECCV 2014* (2014), Vol. 8689, pp. 818–833.
 42. R. Kotikalapudi, “Keras-vis,” <https://github.com/raghakot/keras-vis>.
 43. N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, “Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?” *IEEE Trans. Med. Imaging* **35**(5), 1299–1312 (2016).

1. Introduction

Deep learning methods, driven by large data sets and powerful and sometimes cloud-based computational hardware [1], are transforming computer-aided diagnostics in medicine. Deep learning has been applied to study diseases in the brain [2], skull [3], skin [4], lungs [5], blood vessels [6,7], and eye [8]. Convolutional neural networks have achieved state-of-the-art results on a variety of computer vision tasks and have created new opportunities for applying machine learning to interpret medical image data. The demonstration that convolutional neural networks can classify dermatoscopy images of skin cancer lesions at the level of trained dermatologists [4] is an example of such an application. As new modes of medical imaging become available, it will be important to explore the integration of convolutional neural networks with imaging systems to improve detection of disease.

Convolutional neural networks (CNNs) are a class of deep learning models specialized for computer vision. CNNs are typically made up of many convolution, pooling, and fully-connected layers. A convolution layer convolves the output from the previous layer with a set of learnable filters and then applies an element-wise non-linear activation function, typically $\text{ReLU}(x) = \max(0, x)$ where ReLU stands for Rectified Linear Units. The key property of the convolution layer is that it encodes local connectivity (because the filter size is small) and weight sharing (we use the same filter and hence the same weights at each position in the image), modeling assumptions which are well-suited to images. The convolution layer can thus be interpreted as a set of local feature detectors sliding throughout the image. The pooling layer downsamples along the spatial dimensions by partitioning the output of the previous layer into non-overlapping regions and pooling the values in each region into one output (typically by taking the maximum), thereby locally aggregating information. A stack of convolution and pooling layers allows the network to learn a hierarchy of representations, high-level features from low-level pixels. The fully-connected layers at the end of the network allow us to aggregate all the information that we've learned at the different spatial positions into one final prediction.

In the field of computer vision, convolutional neural networks (CNNs) have achieved state of the art results on many challenging vision tasks related to analyzing and understanding images. Since Krizhevsky et al [9] achieved ground-breaking results on image classification by using CNNs (achieving an error rate of 16.4% compared to the second-place error rate of 26.2% in the 2012 ImageNet competition), CNNs have set new benchmarks in many other computer vision tasks, such as object detection [10], semantic segmentation, image captioning, and image generation. More recently, state of the art computer vision performance has been achieved through more efficient and deeper CNN architectures, as exemplified by Inception [11] and residual networks [12], winners in the 2014 and 2015 ImageNet competitions, respectively. In the case of image classification, one key feature of CNNs is that they can be fully trained in an end-to-end manner, learning the mapping directly from the pixels in the input image to the target output (class labels) from our training data.

Hence, CNNs do not require any kind of hand-crafted rules or feature engineering based on domain expertise. These attributes make CNNs particularly apropos for medical imaging applications.

Optical coherence tomography (OCT) is becoming more widely used to study the auditory portion of the inner ear, the cochlea (Fig. 1), in animal models [13–18]. One goal is to use OCT for non-invasive, direct detection of structural changes in the cochlea that are associated with hearing loss. For example, Meniere's disease is a syndrome of episodic vertigo, fluctuating hearing loss, roaring tinnitus, and aural pressure, that is associated with a specific histological finding called endolymphatic hydrops [19]. Endolymphatic hydrops is the excess accumulation of endolymph fluid within the scala media, a compartment inside the cochlea (Fig. 1(b), 1(c)). While the etiology of endolymphatic hydrops remains poorly understood [20], it can be detected *in vivo* using OCT [21] (Fig. 1(b)) or MRI [22,23], or post-mortem using fixed, sectioned tissue (Fig. 1(c)) [24]. Typically, these imaging techniques provide a cross-section of one or more of the cochlear turns. Subjectively, a cochlear histopathologist can assess the presence of endolymphatic hydrops with relative ease by examining the distension of Reissner's membrane [24], the membrane that separates the scala media from the scala vestibuli.

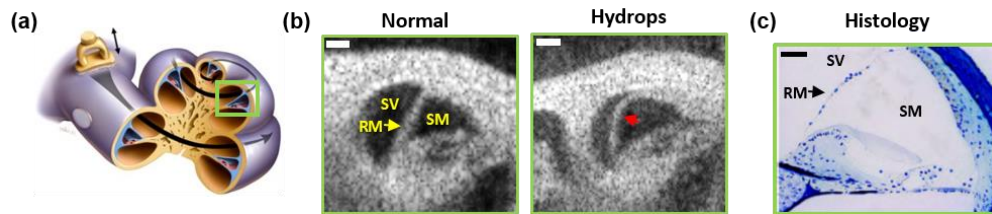


Fig. 1. (a) Schematic of the cochlea. (b) Representative OCT cross-sections of the cochlea in a healthy mouse (*normal*) and a blast-exposed mouse with endolymphatic hydrops (*hydrops*). The red arrow indicates the deformity of Reissner's membrane, the hallmark of endolymphatic hydrops, in the blast-exposed mouse. Scale bars 100 μm . (c) Plastic-embedded section of the upper basal turn in the cochlea of a mouse (adapted from [25]). Scale bar 50 μm . RM, Reissner's membrane; SM, scala media; SV, scala vestibuli.

Objectively, however, establishing the presence or absence of endolymphatic hydrops is not straightforward. A common approach is to measure features based on expert understanding of the disease, such as the deformation of Reissner's membrane [26], the ratio of the area of the scala media to the area of the scala vestibuli [27], and the endolymph volume [20]. An obvious drawback of using expert derived features is that our understanding of the pathology may be incomplete, hence many important features could be omitted by the human designers. Likewise, even well-chosen expert derived features may not span the entire feature space that describes the pathology. Developing software to automatically measure human-designed features is also challenging [28]. A significant fraction of images may be effectively un-analyzable by computer-based methods for human-designed feature extraction, as demonstrated by a recent computer-aided approach to detect endolymphatic hydrops using a measure of Reissner's membrane distension [29].

Here we report an automated approach that uses a CNN architecture, called ELHnet (which stands for EndoLymphatic Hydrops network), to overcome these limitations. We developed ELHnet specifically to classify endolymphatic hydrops in mice undergoing cochlear imaging using OCT. We show that this technique provides accurate and reliable classification of endolymphatic hydrops non-invasively, and has a much higher success rate for analyzing images compared with previous approaches. Thus, this work provides another example where using state-of-the-art convolutional neural networks can provide medical diagnostic capabilities without a priori knowledge of the disease changes, and contributes to the growing literature on the application of CNNs to medical OCT images [30–32].

2. Methods

2.1 Data set

We constructed a data set of 2196 cross-sectional OCT images of cochleae from 54 mice (Table 1). These images were obtained in live, anesthetized mice during biological experiments [15,33]. Briefly, mice were anesthetized with ketamine/xylazine, fixed to a head post, and underwent surgery to open the left middle ear bulla. We then imaged the cochlea using OCT without violating the otic capsule bone. Our 1300 nm swept-source imaging setup has been fully described previously [15]. In order to induce endolymphatic hydrops in mice, nine mice were exposed to a single blast pressure wave with peak pressure of 130 kPa [25] and eighteen mice were exposed to band-passed white noise (8-16 kHz) at 100 dB sound pressure level (SPL) for two hours [21] (a full manuscript detailing the biological features of this phenomenon is currently in preparation). The remaining 27 mice served as normal controls because they received no blast exposure and so did not have endolymphatic hydrops. All protocols were approved by the Institutional Animal Care and Use Committee at Stanford University. Ground truth labels of endolymphatic hydrops were determined visually by a trained observer (J.K.) without blinding during image acquisition, and later reviewed by a second observer (G.S.L.) during image preparation. If the observers reached different determinations of endolymphatic hydrops, then the image was excluded. When multiple cross-sectional images were obtained in the same cochlea, the entire set of cross-sections were viewed together to determine the ground truth label of the cochlea and its cross-sections.

Table 1. Number of mice and images in the training, validation, and test data sets.

Class	Training		Validation		Test	
	Mice	Images	Mice	Images	Mice	Images
Control	6	1015 (60900)	2	260 (2600)	19	19
Hydrops	7	761 (57836)	2	123 (1230)	18	18

The numbers of computationally-augmented images in the training and validation data sets are indicated in parentheses.

In 17 mice, cochleae were imaged from multiple angles, positions, and depths to capture hundreds of cross-sectional images per cochlea. The remaining 37 mice used in the test group had their cochleae imaged only once. The optical resolution of the imaging system was 9.8 μm lateral and 15 μm axial measured in air [15]. We collected images using sampling of 7.5 μm laterally. Bilinear interpolation was used in the axial direction to interpolate from 15 μm to 7.5 μm . Images were cropped to 128 x 128 pixel regions focused on the apical turn of the cochlea where the endolymphatic space is visible.

The data set is split into three subsets: training, validation, and test subsets (Table 1). The training subset contains 1776 images from 13 mice, and was used to train several candidate convolutional neural network models (i.e. fit the model weights using back-propagation) with different model architectures and hyper-parameters. The validation subset contains 383 images from 4 mice, and was used to choose the trained CNN model that generalized best to held-out data, thereby avoiding overfitting of the model to the training data. The test subset contains 37 images from 37 mice, and was used in the final evaluation of the best CNN model to assess its generalization error for new data (i.e. data not previously used in training and validating the model).

2.2 Data augmentation

Training CNN architectures requires substantial computational power and large amounts of labeled data. Labeled data is particularly difficult to obtain in medical research due to regulatory restrictions and cost of acquiring images. For this work, labeled data was limited because OCT imaging of the cochlea is a relatively new technique and our experiments required the use of live mice. Our data set size was small—2196 labeled images, compared

with 14,197,122 labeled images in ImageNet used for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [34].

To overcome this challenge, we computationally augmented the number of images in our training and validation data sets using classification-invariant transformations, such as rotation, translation, and zooming. Data augmentation was performed on training and validation images to simulate the effects of natural perturbations in the image acquisition process. We generated augmented images by rotating the original 128x128 pixel image by a randomly chosen angle between -20 and 20 degrees, shrinking the image by a randomly chosen factor between 1 and 1.33, and cropping a randomly chosen 64x64 pixel patch within the central 96x96 pixel region of the image (Fig. 2). The purpose of cropping from a randomly chosen center was to simulate image translation. Though cropping necessitates loss of image information that may be relevant for classification of endolymphatic hydrops, we believe this to be unlikely given that the loss of information mainly involves the periphery of pre-augmented images. Constraining the cropping to the central 96x96 pixel region ensured that Reissner's membrane would be included in the augmented image, because Reissner's membrane was already located near the center of the pre-augmented image.

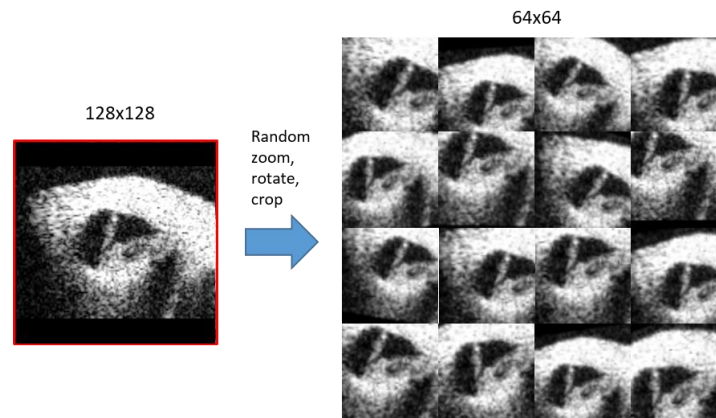


Fig. 2. Process of data augmentation. The original image (left) is rotated by an angle between -20 and 20 degrees, shrunk by a factor between 1 and 1.33, and cropped to a 64x64 pixel region around a randomly chosen center.

The final number of augmented image patches in the training and validation data subsets are shown in Table 1 (*numbers in parentheses*). Specifically, 60 augmented images were generated per training control image and 76 augmented images were generated per training endolymphatic hydrops image; this produced an approximately equal number of augmented images of control and hydrops cochlea for training. In the validation subset, 10 augmented images were generated per validation image.

2.3 ELHnet architecture

The proposed ELHnet architecture is shown in Fig. 3 and detailed in Table 2. The architecture is based on the VGG16 architecture [35], which contains 16 layers (13 convolution layers and 3 fully-connected (“dense”) layers). The ELHnet architecture contains 8 layers, including 6 convolution layers and 2 fully-connected layers. The input to the ELHnet architecture is a two-dimensional 64 x 64 pixel image. We chose this input image size because its dimensions are powers of two, allowing the input image to be processed by multiple pooling layers, which reduce the height and width by half each time as shown in Fig. 3(a). The output is the predicted probability that the input image belongs to the endolymphatic hydrops class. The number of layers in ELHnet balances the classification power of a deep architecture (as

demonstrated by VGG16, the runner-up in ILSVRC 2014) and the reduced overfitting from a shallower architecture on our small data set.

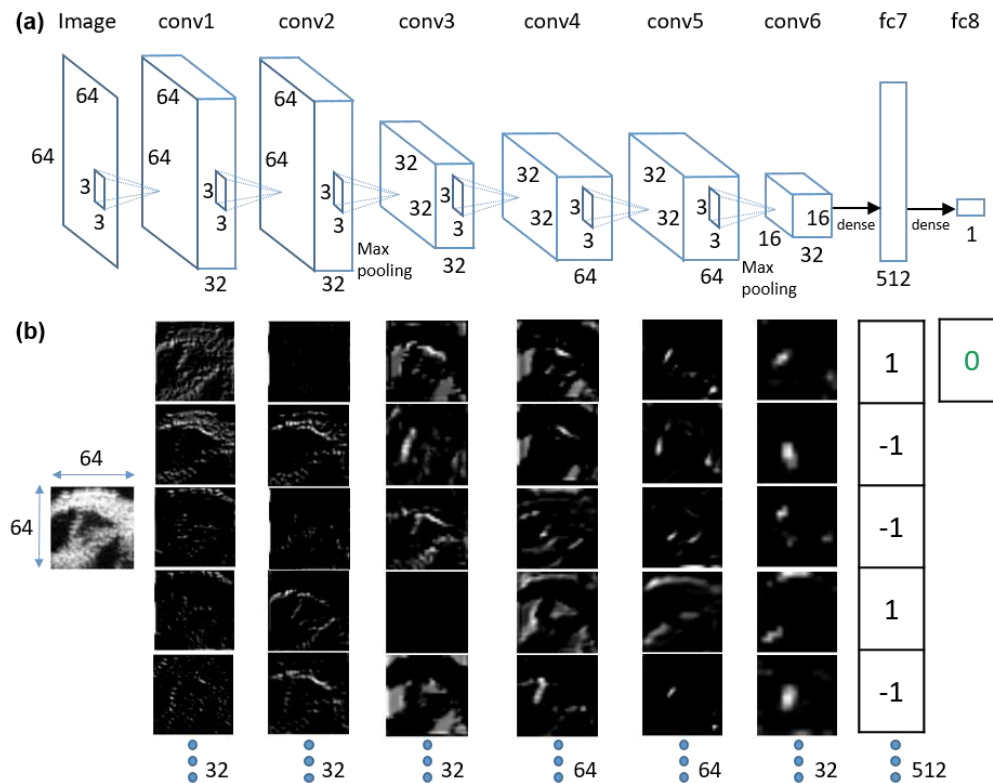


Fig. 3. ELHnet architecture. (a) The layers of ELHnet are illustrated schematically. Schematic design adapted from reference [36]. (b) The first five activations of the output of each layer are visualized, given the example image on the left as the input, resulting in a prediction of no endolymphatic hydrops (*green number*) from the final (fully connected) output layer (fc8). The final output layer (fc8) outputs a score between 0 and 1, which can be interpreted as the model's predicted probability that the input image shows endolymphatic hydrops. The penultimate layer (fc7) tends to output values of either 1 or -1 because of the hyperbolic tangent activation function used for that layer (Table 2), which compresses positive and negative values of large magnitude to 1 and -1, respectively.

Table 2. ELHnet architecture.

Layer	Type	Input	Kernel	Filters	Stride	Pad	Activation	Output
data	Input	1x64x64	N/A	N/A	N/A	N/A	N/A	1x64x64
conv1	Convolution	1x64x64	3x3	32	1	1	ReLU	32x64x64
conv2	Convolution	32x64x64	3x3	32	1	1	ReLU	32x64x64
pool2	Max pooling	32x64x64	2x2	–	2	0	–	32x32x32
conv3 ^a	Convolution	32x32x32	3x3	32	1	1	ReLU	32x32x32
conv4	Convolution	32x32x32	3x3	64	1	1	ReLU	64x32x32
conv5	Convolution	64x32x32	3x3	64	1	1	ReLU	64x32x32
pool5	Max pooling	64x32x32	2x2	–	2	0	–	64x16x16
conv6 ^a	Convolution	64x16x16	3x3	32	1	1	ReLU	32x16x16
fc7 ^b	Fully connected	32x16x16	16x16	512	1	0	tanh	512x1
fc8 ^c	Fully connected	512x1	1x1	1	1	0	sigmoid	1x1

^aTrained using dropout rate of 25%^bTrained using dropout rate of 50%^cTrained using L2-regularization

Convolution layers convolve the output from the previous layer with a set of learnable filters and then applies an element-wise non-linear activation function. The activation functions used were rectified linear unit, $\text{ReLU}(x) = \max(0, x)$; hyperbolic tangent, $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$; and sigmoid function, $s(x) = \frac{1}{1 + e^{-x}}$. Pooling layers

downsample the output of the previous layer along the spatial dimensions by taking the maximum value in each 2x2 region. Fully-connected layers are regular (non-convolutional) dense neural network layers, which aggregate all of the information from the different spatial positions into one final prediction at the end of the network. All weights include a bias term. Weight initialization used the standard Gaussian distribution, which was found to perform better than Glorot normal initialization. Batch normalization was not used. Dropout and L2-regularization were used in the layers indicated for training the model weights.

2.4 Training

The network was trained with the binary cross-entropy loss function. The cross-entropy loss function was motivated by the fact that ELHnet uses a sigmoid activation function in the final output layer, producing an output between 0 and 1 (*fc8* in Table 2). For N images in a batch, the complete loss function L , as a function of the CNN parameters W , of the ELHnet model is defined as:

$$L(W) = -\frac{1}{N} \sum_{i=1}^N \left[y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \right] \quad (1)$$

where $i \in \{1, \dots, N\}$ is the image index, $y^{(i)} \in \{0, 1\}$ is the ground truth label of endolymphatic hydrops in image i , and $\hat{y}^{(i)} = f(x^{(i)}; W) \in (0, 1)$ is the output of the CNN given input $x^{(i)}$ (where $x^{(i)}$ is image i) and weights (parameters) W . $\hat{y}^{(i)}$ can be interpreted as the model's predicted probability that input i belongs to class 1 (i.e. demonstrates endolymphatic hydrops).

Weights of layers were initialized using a standard normal distribution. The network was trained for 50 epochs in batches of 32 images, with early stopping after four epochs without improvement in the loss function. The learning rate was initialized to different values, listed in Table 3, and updated with Adam, a popular stochastic gradient-based optimizer with adaptive learning rates, using all other update parameters following the original paper [37]. Optimal learning rates were determined by trial and error, and are bolded in Table 3. The

training process was carried out using the Theano framework [38] and Keras library [39] in Python 2.7. The networks were trained using an NVIDIA GRID (GK104 “Kepler”) 4GB GPU graphics card on Amazon Web Services Elastic Compute Cloud. The training process took ~11 seconds for an iteration of ~1652 images.

Table 3. Learning rates that were explored during training.

	Learning rate	Accuracy	Sensitivity	Specificity
ELHnet	7.7×10^{-4}	0.9860	0.975	0.9945
	7.8×10^{-4}	0.9906	0.990	0.9910
	7.9×10^{-4}	0.9864	0.987	0.986
	8.0×10^{-4}	0.9768	0.985	0.988
	8.1×10^{-4}	0.9753	0.970	0.980
	9.0×10^{-4}	0.9807	0.975	0.985
VGG-pretrained	7.9×10^{-4}	0.922	0.8855	0.951324
	8.0×10^{-4}	0.9229	0.946	0.904984
	8.1×10^{-4}	0.923	0.936	0.913162
	8.2×10^{-4}	0.924	0.9115	0.933022
	8.3×10^{-4}	0.9175	0.948	0.893692
VGG-finetuned	1×10^{-5}	0.958187	0.9585	0.957944
	1×10^{-4}	0.963004	0.966	0.96067
	1×10^{-3}	0.562172	0	1

Accuracy, sensitivity, and specificity are given for validation performance. The best learning rate, in bold, was used for final training of ELHnet before evaluation on the test data set. ELHnet was trained without using transfer learning. VGG-pretrained used the output of the last convolutional layer of VGGnet as bottleneck features and trained a fully-connected network on top, consisting of three densely connected layers, using the extracted bottleneck features. VGG-finetuned used the VGGnet model weights as initializations to fine-tune the last convolutional block (three convolutional layers) and fully-connected layers (two densely connected layers) using our domain-specific training data set.

To reduce overfitting of the model to training data, we evaluated the trained model on validation data after each epoch, and saved the model weights that maximized validation accuracy over all the model weights that were explored during training. We employed dropout [40] in the third and sixth convolutional layers and first fully connected layer (Table 2) to reducing overfitting during training, as well as L2-regularization of matrix weights in the last fully connected layer (Table 2).

2.5 Evaluation

After training and validation, we evaluated ELHnet on test data and reported the test performance without further modification to the model weights and parameters. To evaluate test data, we wrote code in Python 3.5 that applied the trained ELHnet model to classify new images using a portable and efficient algorithm for matrix multiplication. We wrote code in Python 3.5 so that we would be able to integrate it into our existing OCT software, written in Python 3.5. The computation time for evaluating one image using the trained ELHnet classifier took about 50 to 80 msec.

During evaluation, the user manually selected the center of the region of interest (ROI) in the original image for classification (Fig. 4). The image was then cropped to a 64 x 64 pixel ROI centered on the selected point, and the ROI was input to ELHnet for classification. If the selected center was less than 32 pixels from the image edge, then the image was zero-padded along the image edge to allow cropping of the ROI beyond the image border. ELHnet outputs

the classification score, $\hat{y}^{(i)}$ in Eq. (1), which is between 0 and 1. This is the same output as from the final fully-connected layer, $fc8$, of ELHnet described in Table 2 and Fig. 3. Scores can be interpreted as the model's predicted probability that the input image belongs to the endolymphatic hydrops class. The final classification is determined to be endolymphatic hydrops if the score is greater than 0.5 and non-hydrops otherwise.

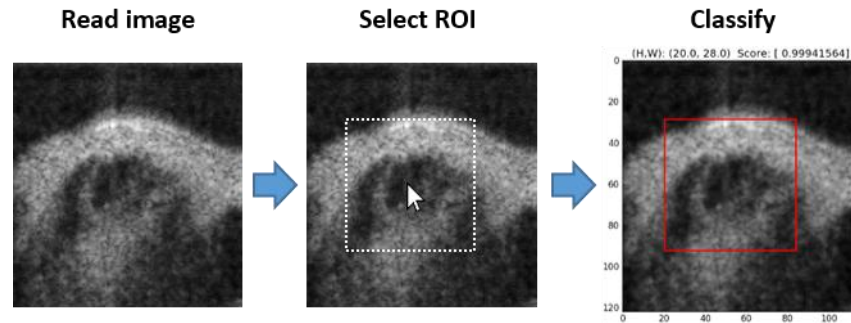


Fig. 4. Steps for classifying images using ELHnet. The uncropped, original image is shown, and a region of interest (ROI) is selected by the user applying ELHnet. The user manually clicks on the center of the desired ROI, whose size is 64x64 pixels. Next, the ROI is input to the ELHnet model for classification. The classification result is shown as a red or green box around the ROI. Green box indicates the predicted class is non-hydrops, and red box indicates the predicted class is endolymphatic hydrops. In this example, the red color of the box indicates a classification of endolymphatic hydrops.

As part of our assessment of ELHnet, we also classified endolymphatic hydrops in test images using an existing computer-aided detection (CAD) approach [29]. The CAD approach applies image processing techniques to measure the perpendicular displacement of Reissner's membrane, i.e. the distance of the midpoint of Reissner's membrane from its normal position on a straight line between the spiral limbus and spiral ligament. Images with displacement measurements above a previously established cut-off value of 7.36 μm [29] were classified as endolymphatic hydrops.

2.6 Visualization

To visualize areas of the input image important for ELHnet classification, we generated rectified saliency maps [41] using an existing Keras visualization package [42]. These maps show the degree to which small changes in individual image pixels influence the prediction of the network (the output from the final fully-connected layer, $fc8$). We displayed the saliency maps in pseudocolor overlaid over the original images in grayscale.

3. Results

3.1 Endolymphatic hydrops classification

We evaluated ELHnet on the validation data set during training and on the test data set after training was completed. The learning curves for the loss function and training and validation accuracies are shown in Fig. 5. Evaluating the final model on the validation set correctly classified all but seven of the augmented validation images (3823/3830 augmented images from four mice). The seven errors were images of endolymphatic hydrops that were misclassified as non-hydrops. These false negatives are shown in Fig. 6, along with a sample of ten true positives and ten true negatives for comparison. For testing, ELHnet correctly classified all 19 non-endolymphatic hydrops images and 15 out of 18 endolymphatic hydrops images (34/37 images from 37 mice). All test images are shown in Fig. 7. Table 4 summarizes the validation and test results. Overall, ELHnet demonstrated 100% specificity in validation and test performance, and 99.4% sensitivity in validation performance and 83.3%

sensitivity in test performance. These results suggest that the training/validation process developed ELHnet to prioritize specificity over sensitivity. Receiver operating characteristic (ROC) analysis of ELHnet demonstrated an area under curve (AUC) of 1.000 for validation data and 0.965 for test data (Fig. 8), suggesting that ELHnet allows sensitive and specific classification of endolymphatic hydrops at multiple decision cutoffs of output classification scores (i.e. besides the cutoff we used of 0.5).

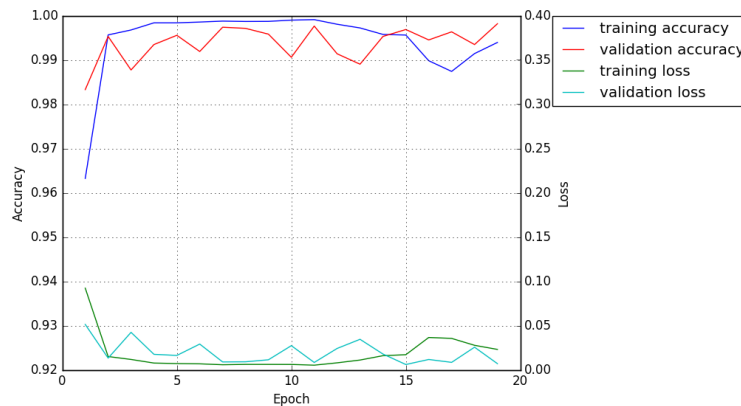


Fig. 5. Learning curves for the loss function and training and validation accuracies during training of ELHnet.

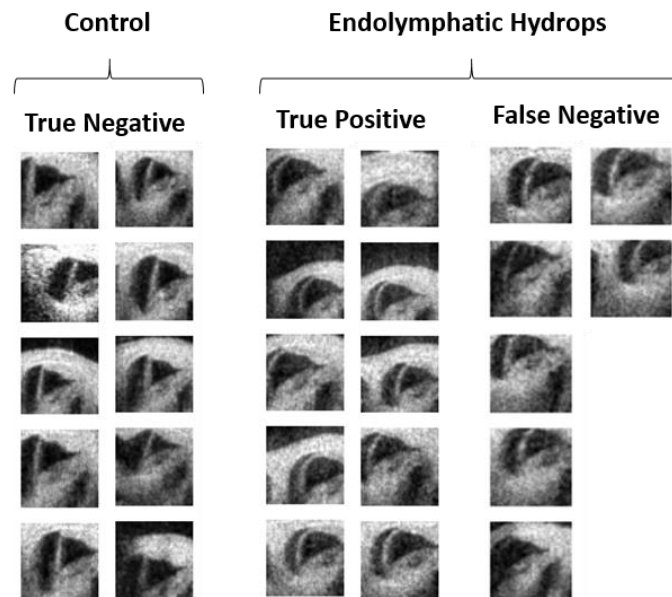


Fig. 6. Validation results for ELHnet classification of the validation data set. Classification was performed using augmented images from the validation data set, i.e. zoomed, rotated, and cropped patches of 64x64 pixels from the original images of 128x128 pixels. All validation images were correctly classified except for seven augmented images of endolymphatic hydrops, which were misclassified as false negatives, as shown. A representative sample of ten true negative and ten true positive augmented images are also shown. Note: a classification result of endolymphatic hydrops was considered “positive” and a classification result of non-hydrops in a control mouse was considered “negative”. Therefore, “true negative” indicates a classification of non-hydrops in a control mouse, “false negative” indicates a classification of non-hydrops in a mouse with endolymphatic hydrops, etc.

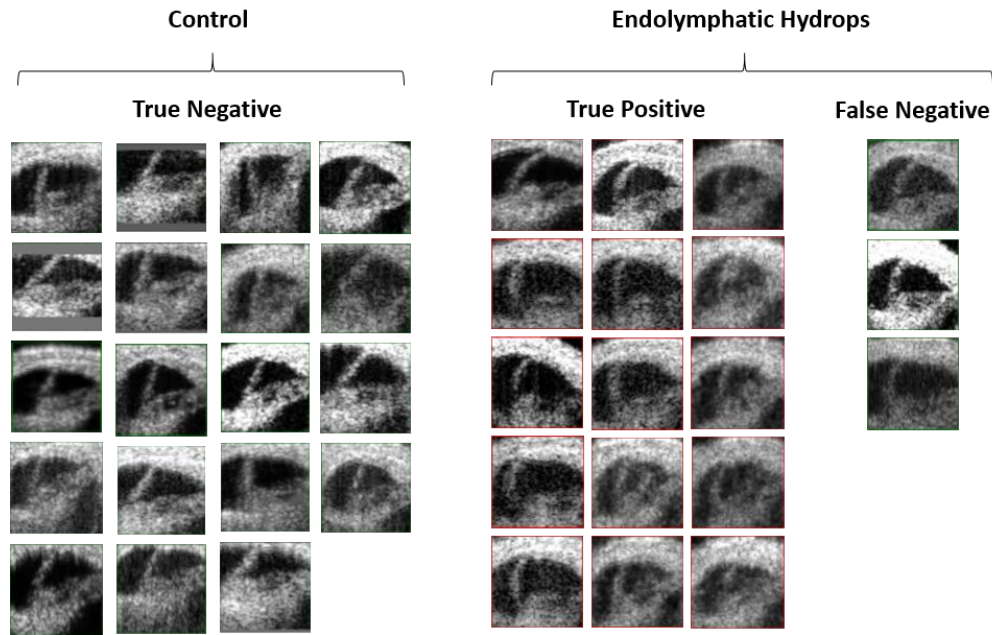


Fig. 7. Test results for ELHnet classification of the test data set. The test data set comprised of previously unseen OCT images, each taken in a new cochlea. Cochleae with ground truth labels of non-endolymphatic hydrops are labeled as control; cochleae with ground truth labels of endolymphatic hydrops are labeled as such. ELHnet correctly classified all 19 control images and 15 of the 18 endolymphatic hydrops images. For clarity, only the region of interest that was used for classification by ELHnet is shown. Note: a classification result of endolymphatic hydrops was considered “positive” and a classification result of non-hydrops was considered “negative”. Therefore, “true negative” indicates a classification of non-hydrops in a control mouse, “false negative” indicates a classification of non-hydrops in a mouse with endolymphatic hydrops, etc.

Table 4. Validation and test performance of ELHnet.

Model	Validation (64x64 patches)			Test		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
ELHnet	0.998 (3823/3830)	0.994 (1223/1230)	1 (2600/2600)	0.919 (34/37)	0.833 (15/18)	1 (19/19)

Validation images (128 x 128 pixels) were augmented to create 10 patches per image, and the augmented patches (64 x 64 pixels) were used for validation. All 2600 validation patches corresponding to the control group were correctly classified, and all but 7 of the 1230 validation patches corresponding to endolymphatic hydrops were correctly classified. When the scores of the 10 patches for each image were averaged to obtain a final prediction for the image, all 123 validation images corresponding to endolymphatic hydrops were correctly classified. For test images, the user manually selected the region of interest (64 x 64 pixels) in the original image for classification.

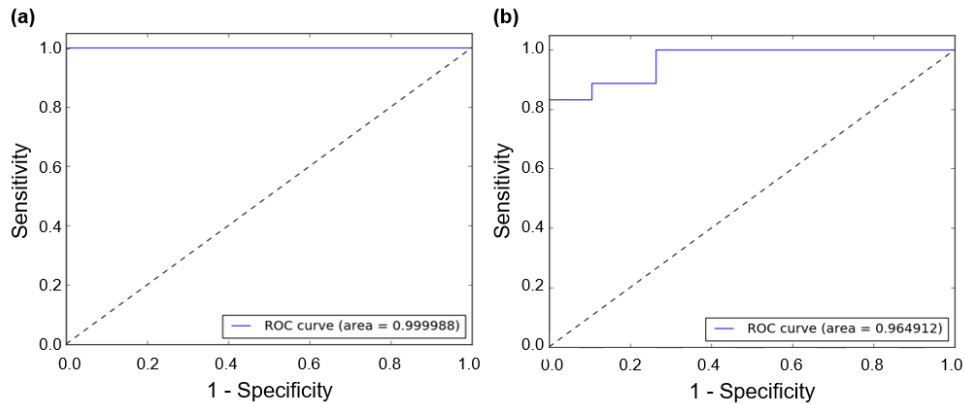


Fig. 8. Receiver operating characteristic curves for ELHnet based on its classification scores for (a) the validation data set and (b) the test data set.

3.2 Performance of ELHnet versus models trained using transfer learning from VGGnet

The validation performances of models trained using the ELHnet architecture were compared with the validation performances of models trained using transfer learning from the pre-existing network VGGnet [35]. Transfer learning has been shown to be an effective method to train deep networks using small data sets common to medical research [43]. We compared ELHnet models with VGG-pretrained and VGG-finetuned models (Table 3). ELHnet was trained (without transfer learning) on our set of medical images with randomly initialized weights. VGG-pretrained uses the pre-existing VGGnet as a fixed convolutional feature extractor and, given the convolutional features as input, only trains the fully connected layers with randomly initialized weights on our set of medical images. VGG-finetuned uses the VGGnet weights as initialization but then has the last block of convolutional layers and the fully-connected layers fine-tuned. Fine-tuned means that the weights were initialized with the pre-existing VGGnet weights and then trained on our set of medical images. Overall, ELHnet demonstrated higher sensitivity and specificity in validation performance than VGG-pretrained and VGG-finetuned.

3.3 Performance of ELHnet versus an existing computer-aided detection approach

To further assess ELHnet, we compared its classification performance on the test data set with that of an existing computer-aided detection (CAD) approach [29]. Similar to ELHnet (and unlike VGGnet), the CAD approach was specifically designed to classify endolymphatic hydrops in cochlear OCT images, thus providing a relevant baseline for comparison. The CAD approach correctly classified 10 of 16 analyzed control images and 14 of 15 analyzed hydrops images, and was unable to analyze 3 control images and 3 hydrops images (Fig. 9). Notably, five of the six false positive images were misclassified due to errors in segmentation of Reissner's membrane that resulted in erroneously large measurements of perpendicular displacement. Compared with ELHnet, the CAD approach showed about the same sensitivity (93.3% vs 91.9%) but a substantially lower specificity (62.5% vs 83.3%). Moreover, ELHnet is better than the existing CAD approach because it has the ability to analyze all input images instead of only a fraction of them (50-87% in previous work [29]).

We used the previously established cut-off value of $7.36 \mu\text{m}$ to classify endolymphatic hydrops using CAD measurements of perpendicular displacement [29]. To assess sensitivity and specificity at other cut-off values, we also performed ROC analysis and found the AUC to be 0.775 for the CAD approach (Fig. 10). This was less than the AUC of 0.965 for ELHnet (Fig. 8).

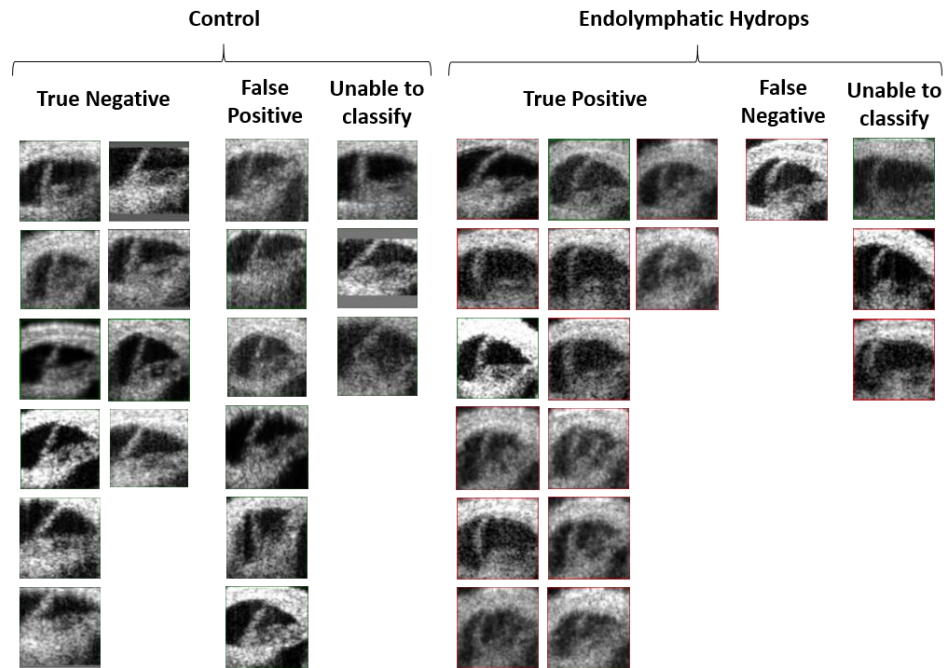


Fig. 9. Test results for the computer-aided detection (CAD) approach [29] classification of the test data set. This is the same test data set used to evaluate the test performance of ELHnet. The CAD approach correctly classified 10 of the 16 analyzed control images and 14 of the 15 analyzed endolymphatic hydrops images. The CAD approach was unable to analyze (and therefore classify) 3 control images and 3 hydrops images.

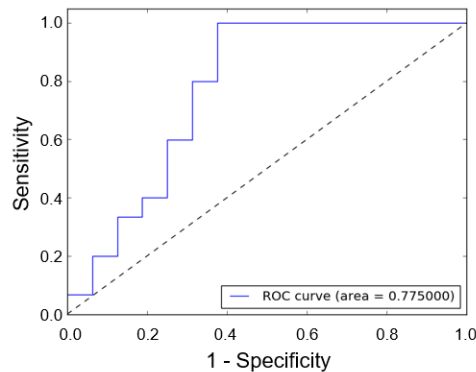


Fig. 10. Receiver operating characteristic curve for the computer-aided detection approach [29] to classifying endolymphatic hydrops in the test data set.

3.4 Visualization of ELHnet using saliency maps

To visualize areas of the image ELHnet is dependent on for classification, we generated saliency maps [41] for correctly classified, pre-augmented 64x64 training images (Fig. 11(a)). We observed salient pixels along Reissner's membrane, as well as along the tectorial membrane and otic capsule near Reissner's membrane's attachment to the spiral ligament. In the three test images that ELHnet misclassified as false negatives, the salient pixels were localized to the proximal upper portion of Reissner's membrane and approximated a short straight line (Fig. 11(b)). Overall, these visualizations suggest that Reissner's membrane is an important learned feature for ELHnet classification.

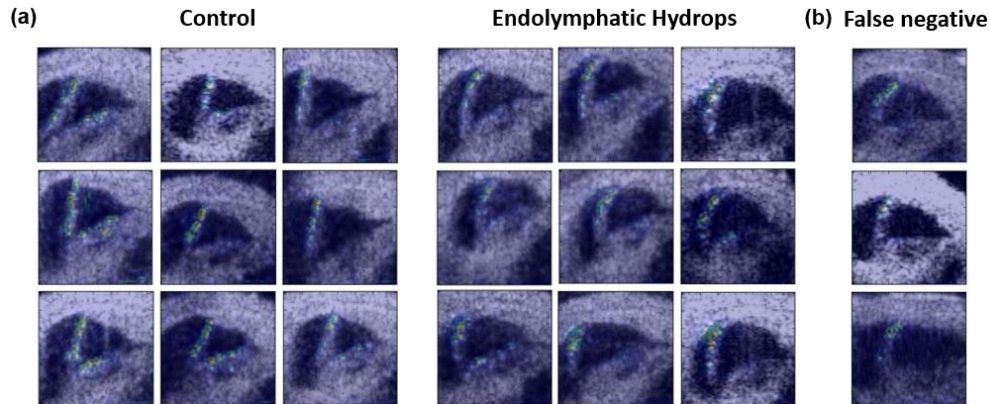


Fig. 11. Saliency maps for (a) control and endolymphatic hydrops training images correctly classified by ELHnet, and (b) the three test images of endolymphatic hydrops misclassified by ELHnet. Saliency maps are overlaid in pseudocolor over the original images in grayscale.

4. Discussion

In this paper, we address the problem of endolymphatic hydrops classification in cochlear OCT imaging using a convolutional neural network (CNN) architecture called ELHnet. We demonstrate using experimental data that this approach can automatically analyze all input images and classify them with high specificity and moderate-to-high sensitivity. The procedure for ELHnet classification is automated after the user clicks on the scala media in the OCT image.

To train and evaluate ELHnet, we developed a data set containing 2196 OCT images of cochleae from 54 mice. This is a larger number of mice used compared with previous work in automated endolymphatic hydrops classification [29]. This is also the first demonstration of a convolutional neural network approach to endolymphatic hydrops classification.

During training, we observed significant improvements in the learning curves after one epoch (Fig. 5). This may be influenced by data augmentation, which increased the number of training images per epoch by over 60 fold (Table 1). Learning curves plotted per batch of 32 images, instead of per epoch, could allow improved visualization of the learning process during the first epoch. We note that the validation accuracy is initially greater than the training accuracy after the first epoch, which might be because the validation images were easier to classify than the training images due to small sample variation.

We sought to identify image features that could explain the misclassification of false negative images by ELHnet. The saliency maps suggest that Reissner's membrane is an important learned feature (Fig. 11), but it is not clear what would lead ELHnet to misinterpret this learned feature in the images that were misclassified. Based on a qualitative assessment, false negative images in the validation data set tended to have a higher-zoom level and a broader arc of curvature to Reissner's membrane compared with true positive images (Figs. 6 and 7). Frankly, it is also difficult for a human observer to distinguish whether some of these borderline images have increased or normal endolymph volume.

It is possible that misclassification owes in part to overfitting, i.e. classification of images using non-intuitive image features emerging from sample noise rather than true underlying patterns in the training data. The risk of overfitting is increased when the training sample size is small [40]. Lack of generalizability also appears to be an issue, given the higher sensitivity of ELHnet classification observed for validation data (99.4% sensitivity) compared with for test data (83.3% sensitivity). This may be in part due to the fact that training and validation data were generated using image augmentation whereas test data was not. Systematic differences in the experiments used to acquire the validation and test images (e.g. resulting in

systematic variation in endolymphatic hydrops severity, image clarity, speckle noise, resolution) could also contribute to the validation and test performance discrepancy.

This study has some limitations. It should be noted that the OCT images come from one imaging setup and research team. This may limit variation in images that could owe to differences in animal preparation or imaging setup (e.g. laser power, positioning and angling of the mouse cochlea). Training a CNN on such a data set with limited variation may result in overfitting and reduced generalizability to other data sets. To overcome this challenge, images from other research teams and other OCT systems should be incorporated in future data sets. A second limitation is that this network was trained and tested using only the left mouse cochlea. We imposed this constraint to reduce training time and data storage. A straightforward solution to classifying images of right cochleae is to re-train ELHnet using the horizontally-flipped and non-flipped image of each augmented image. Finally, the performance of ELHnet should be evaluated on a larger number of test images to confirm the network's performance characteristics. The small size of the test data set (37 images total) owes in part to our requirement that each test image represent an entirely new cochlea. The relative lack of available OCT images of endolymphatic hydrops has made acquiring a large data set challenging. Hopefully, as more OCT imaging experiments are performed in the cochlea, it will be possible to train and test CNN models with larger amounts of data to improve upon the performance demonstrated by ELHnet.

We have shown the applicability of ELHnet for endolymphatic hydrops classification. This application may be important for research in endolymphatic hydrops and its relationship to hearing loss, especially as OCT imaging of the cochlea becomes more widely used. If/when OCT imaging is developed for the human cochlea, ELHnet could also aid the detection of endolymphatic hydrops in patients. However, OCT imaging of the human cochlea remains a challenge because of the greater thickness of the cochlear bone. Several unanswered questions remain for automated classification of endolymphatic hydrops, such as how ELHnet and other automated methods compare with single-observer classification of pre-labeled images (by expert reviewers). In future work, we aim to compare ELHnet with other existing methods for classifying endolymphatic hydrops so that researchers will have a better understanding of which method may best suit their needs.

5. Conclusion

We provide, to our knowledge, the first deep learning approach to classifying endolymphatic hydrops using cross-sectional OCT images of cochleae. Classification by ELHnet, our trained convolutional neural network, yielded test performance of 83.3% sensitivity (15/18 images) and 100% specificity (19/19 images). These findings support the role of convolutional neural networks in classifying endolymphatic hydrops for research and future clinical applications.

Funding

National Institutes of Health National Center for Advancing Translational Science Clinical and Translational Science Award (UL1 TR001085); NIH-NIDCD (DC014450, DC013774, DC010363); Stanford Medical Scholars Research Program.

Acknowledgments

The authors would like to thank Dr. James Dewey for help with acquiring data, Darwin Yi for assistance with deep learning software, and the rest of the Oghalai Lab members for helpful discussion. Code in Python is available at GitHub in the "jsol11/ELHnet" repository.

Disclosures

The authors declare that there are no conflicts of interest related to this article.